NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
School of Computer Science and Engineering
College of Engineering

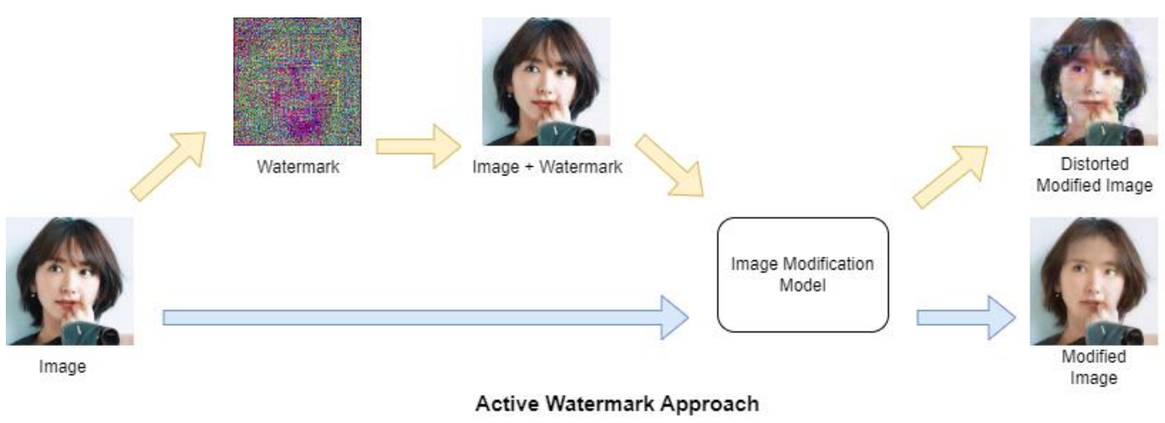
Watermark for combating deepfakes

Improving Watermark Visual Quality

Student: Li Rui
Supervisor: Prof. Lin Weisi, Mr. Hou Jingwen

Introduction

One noticeable approach for combating Deepfake is to apply the model's adversarial noise as a watermark to the image so that when the image is modified, it would be drastically distorted.



Active Watermark Approach

However, current works, e.g., CMUA [1] set the adversarial noise threshold relatively high, making the noise visible on human faces, impairing the image quality. Thus, the goal of this work is to produce universal watermark with improved visual quality, while still maintaining its protection performance.

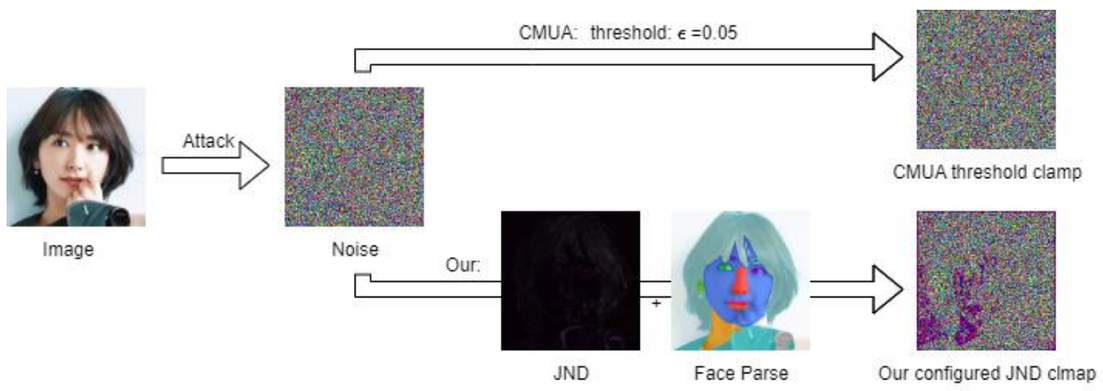
Challenge

There are two main challenge we face:

1. Large parts of an image have a low tolerance for noise. The noise allocated in these areas can easily become visible.
2. To ensure the desired protection performance, certain amount of noise must be allocated.

Method

- Replace adversarial noise threshold with JND to guide noise allocation.
- Segment face into different parts and scale the JND values differently for different parts to increase to total amount of allowed noise.

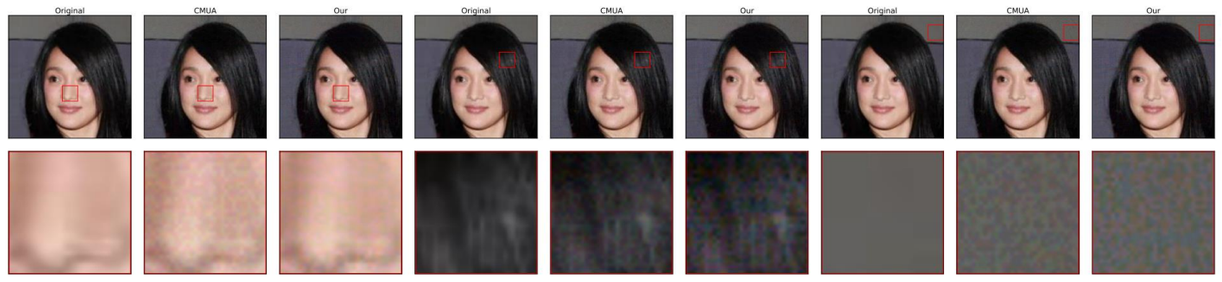


Experiment Results

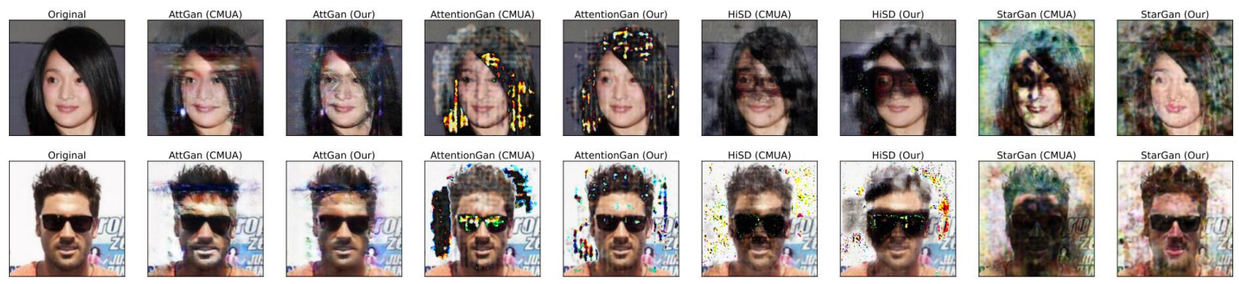
Quantitatively, we successfully achieved improvements both on visual quality and protection performance.

Method \ Results	Protection Performance				Visual Quality	
	Star [2]	Attention [3]	Att [4]	HiSD [5]	SSIM	SSIM Face
CMUA	**1.0**	0.9949	0.8496	0.9869	0.8897	0.9744
Our	0.9976	**0.9998**	**0.8872**	**0.9990**	**0.8920**	**0.9851**

Qualitatively, it achieves better visual effect in face, resulting in better visual quality.



Moreover, although with a little scarifies in the distortion level, our method can deliver the required distortion to differentiate modified images.



Conclusion

Noticing the visual quality defects in the watermarks, we focus on improving visual quality while maintaining protection performance. Experiments show that not only have we managed to achieve better visual effect, but also, we are able to guard more images against modification by exploiting the properties of JND.

References

[1] Hao Huang et al. "Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes". In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. 1. 2022, pp. 989–997.
[2] Yunjey Choi et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation". In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 8789–8797.
[3] Hao Tang et al. "Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks". In: IEEE Transactions on Neural Networks and Learning Systems (2021).
[4] Zhenliang He et al. "Attgan: Facial attribute editing by only changing what you want". In: IEEE transactions on image processing 28.11 (2019), pp. 5464–5478.
[5] Xinyang Li et al. "Image-to-image translation via hierarchical style disentanglement". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 8639–8648.